BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (http://bmjopen.bmj.com).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open

First Author Ethnicity and Gender Predict Publication of BMJ Letters to the Editor

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-037269
Article Type:	Original research
Date Submitted by the Author:	26-Jan-2020
Complete List of Authors:	Zeina, Mohamad; Barts Health NHS Trust Balston, Alfred; Barts Health NHS Trust Banerjee, Amitava; University College London, Farr Institute of Health Informatics Research Woolf, Katherine; University College London, Research Department of Medical Education
Keywords:	Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, EPIDEMIOLOGY, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Health economics < HEALTH SERVICES ADMINISTRATION & MANAGEMENT

SCHOLARONE™
Manuscripts

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

For peer review only

First Author Ethnicity and Gender Predict Publication of BMJ Letters to the Editor

Dr. Mohamad Zeina[1], Dr. Alfred Balston[1], Dr. Amitava Banerjee[2], Dr. Katherine Woolf[3]

[1] Foundation Year 1 Doctor, Barts Health NHS Trust, London, United Kingdom

[2] Associate Professor in Clinical Data Science and Honorary Consultant Cardiologist, Institute of Health Informatics, University College London, London, United Kingdom

[3] Associate Professor in Medical Education, Research Department of Medical Education, University College London Medical School, Royal Free Hospital, London, United Kingdom

Correspondence: Mohamad Zeina

Whipps Cross University Hospital

Barts Health NHS Trust,

London, E11 1NR

Tel: +44 7827353938

Email: mohamad.zeina@nhs.net

Copyright: The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd to permit this article (if accepted) to be published in BMJ editions and any other BMJPGL products and sublicences such use and exploit all subsidiary rights, as set out in our licence.

Competing interest statement: All authors have completed the Unified Competing Interest form and declare:

Dr. Zeina has nothing to disclose.

Dr. Balston has nothing to disclose.

Dr. Banerjee reports personal fees from Boehringer-Ingelheim, personal fees from Astra-Zeneca, personal fees from Novo-Nordisk, personal fees from Pfizer, outside the submitted work; and I am a Trustee of the South Asian Health Foundation, and a member of the Education Committee of the British Cardiovascular Society.

Dr. Woolf reports grants from National Institute for Health Research, non-financial support from Membership of the Royal Colleges of Physicians (UK) examination.

Guarantor: Mohamad Zeina

Contributorship statement:

For peer review only

- Mohamad Zeina: conceptualisation, design and coordination of project, lead in programming and statistical analysis, writing first draft.
- Alfred Balston: contribution to programming, statistical analysis, and writing first draft.
- Amitava Banerjee: aiding in conceptualisation, offering statistical advice, proofreading.
- Katherine Woolf: conceptualisation and design, offering statistical advice, proofreading.

The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted

Transparency declaration: Mohamad Zeina, affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Ethical approval: Ethical approval was not required.

Funding, sponsors and independence: This work is independent from funders, however other funding received by the authors is declared above in the competing interests statement.

Patient and Public Involvement: Patients were not involved in this study.

Dissemination statement: Dissemination not applicable as this study involved no participants or patient organisations.

For peer review only

Abstract

Objectives: To analyse the relationship between first author gender and ethnicity (estimated from first name and surname), and chance of publication of rapid responses in the BMJ. To analyse whether other features of the rapid response account for any gender or ethnic differences, including presence of multiple authors, declaration of conflicts of interests, presence of Twitter handle, word count, reading ease, spelling and grammatical mistakes, and presence of references.

Design: Retrospective observational study.

Setting: Website of the BMJ (BMJ.com)

Participants: Publicly available rapid responses submitted to BMJ.com between 1998 and 2018.

Main outcome measures: Publication of a rapid response as a letter to the editor in the BMJ.

Results: We analysed 113,265 rapid responses, of which 8,415 were published as letters to the editor (7.4%). Statistically significant univariate correlations were found between odds of publication and: first author estimated gender and ethnicity, multiple authors, declaration of conflicts of interest, presence of Twitter handle, word count, reading ease, spelling and grammatical mistakes, and presence of references. Multivariate analysis showed that first author estimated gender and ethnicity predicted publication after taking into account the other factors. Compared to white authors, black authors were 26% less likely to be published (OR 0.74, CI 0.57-0.96), Asian and Pacific Island authors were 46% less likely to be published (OR 0.54, CI 0.49-0.59), and Hispanic authors were 49% less likely to be published (OR 0.51, CI 0.41-0.64). Female authors were 10% less likely to be published (OR 0.90, CI 0.85-0.96) than male authors.

Conclusion: Ethnic and gender differences in rapid response publication remained after accounting for a broad range of features, themselves all predictive of publication. This suggests that the reasons for the differences of these groups lies elsewhere.

Strengths and limitations of this study

- This study utilises corpus of publicly available data to analyse correlations between first author characteristics and chance of publication of letters to the editor in the BMJ.
- Multivariate analysis allowed us to account for a range of other features of submitted letters.
- To our knowledge, this is the largest ever analysis of a scientific corpus that looks at associations with publication rate.
- The nature of this data means that only associations can be inferred, and not causation.
- We highlight automated techniques that scientific journals can use to look for associations between ethnicity and gender in their own publication rates.

Introduction

Commented [MZ1]: We have reworded parts of this to make it easier to read, following a comment from reviewer four.

Much has been written about the "attainment gap" or "differential attainment"; the observation that many fields exhibit discrepancies in achievement based on personal attributes such as gender and ethnicity. In medicine, for example, students from black and minority ethnic (BME) groups achieve poorer marks and are more likely to fail, on average, than their white counterparts (1). As they progress in their careers, BME doctors also more often fail their specialty training exams (1,2), earn a lower average salary than others at the same level of seniority (3), and are less likely to be awarded funding grants (4). There also remain discrepancies in the representation of women in medical leadership and faculty despite a long history of roughly equal proportions of male and female medical students (5,6).

Another specific area where the effects of gender have been studied thoroughly is academic publishing. A survey of 1065 authors from different backgrounds found that women were underrepresented in the scientific literature, along with certain ethnic minorities (7). In a group of high impact medical journals, including the British Medical

Journal (BMJ), the number of articles with female first authors has increased over time; however the gender balance of last authors, who are typically senior researchers or heads of departments, has not followed this trend (8). A larger scale analysis of 1.8 million scholarly articles indexed in JSTOR found female authors are poorly represented in the prestigious first and last author positions in the majority of academic disciplines (9). The same result was observed in an analysis of 21 million articles indexed in Medline which also found BME authors are similarly less likely to be in the last author position even when accounting for seniority. (10).

The cause of these gender and ethnic differences remains the subject of debate. Experimental studies have shown that identical submissions randomly assigned to have a male or female name are ranked differently depending on the gender of the applicant's name, favouring men (11,12). For example, applications for a laboratory manager position that were assigned a male name were rated as being significantly more competent and hireable by faculty members compared to identical applications assigned a female name (11). Similarly, a study presented graduate students with a sample of abstracts from an international conference, where abstracts were randomly shown to have male or female authors. The study found that abstracts presented as having female authors were deemed to have lower "Scientific Quality" (12). Both papers found that the gender of the reviewer did not affect how applicants were rated, and concluded that pervasive gender stereotypes create a subtle but significant bias against women (11,12). Indeed, the bias was reduced when reviewers' attitudes towards gender roles were taken into account, with higher support for gender equality being associated with higher ratings for female authors.

Some argue that gender disparities arise from men and women choosing different career paths (13). For example, women may opt to prioritise flexibility or take time out of their career to have children. However, studies that incorporate these factors into multivariate statistical models fail to fully account for discrepancies in pay (14,15). One such study found a $14,581 yearly salary difference between male and female hospital physicians in the United States which remained after accounting for differences in job satisfaction priorities between genders (16).

The underlying causes of these differences are likely to be complex and multifactorial, but identifying and characterising disparities in new specific situations might hint at potential solutions. These may be broadly applicable, especially because the causative issues are likely to compound each other. For example, lower average pay for women and BME doctors may be partly due to lower chances of scientific publication, especially in a work environment where publication in the scientific literature is important for attaining certain senior academic and leadership positions.

Though many studies mentioned here find group differences based on personal characteristics, they rarely have access to raw data from the journals that would be necessary to quantify publication rate. For example, a finding that women are under-represented in authorship of medical journal papers compared to in the medical workforce is not enough to draw conclusions about discrimination or bias, as it may be due to differences in priorities and the number of submissions sent. For a study to draw meaningful conclusions regarding discrepancies in acceptance rates, it must be able to quantify the percentage of submitted scientific works that are accepted, and this submission data is seldom released by scientific journals.

Letters in the BMJ are derived from rapid responses, which are available online freely and in their entirety, and therefore they may provide a valuable perspective for looking at this issue. Moreover, publication of rapid responses is of importance since letters to the editor carry PubMed identifiers (PMIDs) and thus discrepancies in their publication may have knock-on effects for jobs in academia where PubMed indexed publications play an important role in candidate selection.

We aimed to compare the corpus of available rapid responses with published letters to the editor to look for correlations between ethnicity, gender, and odds of publication.

For peer review only

Methods:

Data acquisition and processing

An automated script was used to download every BMJ.com online rapid response between 25th April 1998 and 23rd March 2018, as well as every letter to the editor that was published in the same timeframe.

To minimise the impact on BMJ servers, webpage requests were only sent every 15 seconds, and each request explicitly stated a full name and contact email address of the researcher carrying out the automated data collection, so that they could easily be contacted if the BMJ wished this collection to stop. Further, we only collected publicly available data that can be accessed without a login to a BMJ account.

Once collected, every available field from the rapid response was extracted. This included: title, title of article being responded to, body of text, first author name, first author title, other authors, date of submission, and presence of Twitter handle. Further processing with software packages mentioned below allowed us to look at a richer set of features including: word count, presence of references, number of references, Flesch reading ease (a measure of complexity of language, with a higher value meaning easier to read), number of spelling and grammatical mistakes, gender of first author, ethnicity of first author, and presence of multiple authors.

The position of the author was extracted by looking for the presence of each of the words "Consultant", "Professor", "Senior" and "Student" in the self-declared occupation field of submitted rapid responses, for example someone who had the word "Consultant" anywhere in their occupation field was classed "Consultant".

We did not expect a linear relationship between publication and word count or Flesch reading ease, because the most successful letters are likely to be long enough to offer a meaningful insight into the topic, but not too long as to be unsuitable for the short letter to the editor format. Thus we created two additional features from these, "Near ideal word count" and "Near ideal Flesch reading ease" to reflect whether a rapid response was within the 50% of rapid responses which are closest in word count and Flesch reading ease to the numbers which have historically been associated with higher rates of publication.

Some rapid responses (<2%) could not be collected automatically due to errors in their formatting which prohibited their automated collection. These rapid responses were omitted from analysis. Regarding collected rapid responses, the absence of data was itself useful information (for example, the absence of 2nd authors was processed as there being no second 2nd authors) and so no analysed data point was considered missing.

As there is a lag between submission of a rapid response and publication of the response as a letter, we excluded all rapid responses that were within 66 days of our data collection window (i.e. submitted after 16th January 2018). This value was based on preliminary analysis that found a vast majority (80%) of letters were published within 66 days of the rapid response submission date.

Matching protocol

Although both rapid responses and published letters are available freely on BMJ.com, they are available on different parts of the website, and the vast majority of published letters do not link to the specific rapid response that was initially submitted. The task of finding out which rapid responses have been accepted is further complicated by the fact that that many editorial changes are made between the submission of a rapid response, and it being printed in the BMJ. Therefore, finding the corresponding rapid response for a letter is not as trivial as looking for a rapid response with identical text content.

To carry out this correspondence, a hierarchical matching protocol was used, which we summarise here. For each published letter, we search the corpus of rapid responses for those by the same first author. To make this possible, author names were standardised by removing middle names or initials. When a first author was only associated with a single rapid response, and a single letter to the editor, these were designated as the same submission. When the

For peer review only

author of a letter to the editor has submitted numerous rapid responses, one was chosen where the first 50 characters had the highest similarity with the letter to the editor.

If no rapid response could be found with the same first author as the published letter to the editor, author name was ignored and rapid responses submitted recently before publication of the letter to the editor were searched for one with the highest similarity in the first 50 characters. A subset of 600 matched rapid responses and letters was checked manually by AJB and MZ and found to be 85.3% accurate.

Classifying ethnicity and gender

Authors of rapid responses are not asked to disclose their ethnicity or gender, and the number of rapid responses involved was too great to individually contact the authors and ask this sensitive information. An automated method was used that could determine ethnicity of a name, for many tens of thousands of names, quickly and with little manual input. This took the shape of a previously published machine learning algorithm, *nameprism.com*, that has been trained to classify ethnicity on 74 million names (17). To the best of our knowledge, it has demonstrated the highest classification accuracy of any publicly available tool for this task. We follow previous medical research that has used name to classify ethnicity (18,19), as well as a validation study that suggested name analysis is accurate enough to be used to aid health research (20).

This ethnicity classification tool was trained on a large, diverse set of names which the authors claim cover 90% of the world's names (17). It was developed in the US and so the six ethnicity categories used are American: White, Black, Asian and Pacific Islander (API), Hispanic, American Indian and Alaska Native (AIAN) and more than two races.

Gender of first author name was determined using a tool called *Gender Guesser* which utilises a database of approximately 40,000 common names and their corresponding gender (21). Names from rapid responses are checked against this database, and placed into one of the following categories: Male, Female, Mostly Male, Mostly Female, Androgynous (equal probabilities of being male or female) and unknown (not in the database).

Statistical analysis

Univariate associations between author and rapid response features, and publication was carried out by calculating chi square and t-test scores. Hierarchical binary logistic regression was used to look at the correlation between ethnicity and publication, taking into account other author and rapid response features.

Software used

Gender of first author names was classified using *Gender Guesser* (21). Ethnicity was classified using *nameprism.com* (17). Flesch reading ease score was calculated using an open source library called *textstat* (22). Spelling and grammatical mistakes were quantified using a tool called *language-check* (23).

All code was written in "Python 3" in the "Jupyter notebook" text editor (24). Data collection used an automated script, utilising the open source Python libraries "Requests" and "BeautifulSoup" (25,26). Further processing and data manipulation used the Python libraries "NumPy", "Pandas" and "SciKit learn" (27–29). Statistical analysis was carried out in the IBM SPSS 25 package and in Python.

Results:

Baseline data

113,265 rapid responses were analysed, of which 8,415 (7.4%) were published as letters to the editor. Of all submitted rapid responses, 83% had first authors with names classed as "white"; 62% of first authors were classed as "male". See table 1 for baseline author and rapid response features.

For peer review only

Table 1: Characteristics of published and unpublished rapid responses submitted to BMJ.com between 25th April 1998 and 23rd March 2018. API and AIAN stand for "Asian and Pacific Islander" and "American Indian and Alaska Native" respectively.
Plus-minus values are means ± SD. Percentages may not sum to 100% due to rounding.

Commented [MZ2]: We have spelled out these abbreviations as suggested by reviewer one.

Characteristic	All submissions (n=113265)	Published (n=8415)	Unpublished (n=104850)
Author gender [number (%)]			
Male	70256 (62.0)	5636 (67.0)	64620 (61.6)
Female	18592 (16.4)	1409 (16.7)	17183 (16.4)
Mostly male	2434 (2.1)	171 (2.0)	2263 (2.2)
Mostly female	1321 (1.2)	98 (1.2)	1223 (1.2)
Androgynous	1021 (0.9)	82 (1.0)	939 (0.9)
Unknown	19641 (17.3)	1019 (12.1)	18622 (17.8)
Author ethnicity [number (%)]			
White	94077 (83.1)	7492 (89.0)	86585 (82.6)
API	15759 (13.9)	726 (8.6)	15033 (14.3)
Hispanic	1903 (1.7)	90 (1.1)	1813 (1.7)
Black	1204 (1.1)	64 (0.8)	1140 (1.1)
AIAN	2 (0.0)	0 (0.0)	2 (0.0)
Unknown	320 (0.3)	43 (0.5)	277 (0.3)
Word count	314 ± 318	410 ± 278	307 ± 319
Flesch reading ease	50.3 ± 16.3	47.6 ± 12.1	50.5 ± 16.5
Has references [number (%)]	40173 (35.5)	4445 (52.8)	35728 (34.1)
Number of references	1.3 ± 3.0	2.2 ± 3.3	1.3 ± 2.9
Author position [number (%)]			
Consultant	16291 (14.4)	1592 (18.9)	14699 (14.0)
Professor	9959 (8.8)	1110 (13.1)	8849 (8.4)
Senior	4491 (4.0)	523 (6.2)	3968 (3.8)
Student	3080 (2.7)	143 (1.7)	2937 (2.8)
Other	79444 (70.1)	5047 (60.0)	74397 (71.0)
Twitter handle present [number (%)]	1868 (0.2)	254 (0.3)	1614 (0.2)
US spelling and grammar errors	27.0 ± 44.6	35.2 ± 39.2	26.3 ± 44.9
UK spelling and grammar errors	8.9 ± 18.0	9.4 ± 15.3	8.9 ± 18.2
Multiple authors [number (%)]	19256 (17.0)	2914 (34.6)	16342 (15.6)
Competing interests declared [number (%)]	6184 (5.5)	924 (11.0)	5260 (5.0)

Univariate analysis

Univariate associations were found between rapid response publication and: presence of references (chi square = 1196.128, df = 1, p < 0.0005), declaration of competing interests (chi square = 536.745, df = 1, p < 0.0005), weekday of submission (chi square = 108.825, df=6, p < 0.0005), first author title containing the word "student" (chi square = 35.748, df = 1, p < 0.0005), first author title containing the word "consultant" (chi square = 151.853, df = 1, p < 0.0005), first author title containing the word "professor" (chi square = 219.259, df = 1, p < 0.0005), first author title containing the word "senior" (chi square = 120.862, df = 1, p < 0.0005), presence of multiple authors (chi square = 2001.860, df = 1, p <0.0005), presence of a twitter handle (chi square = 105.063, df = 1, p < 0.0005), month of submission (chi square 39.581, df = 11, p < 0.0005), word count being close to ideal (chi square 561.003, df = 1, p < 0.0005), Flesch reading ease being close to ideal (chi square = 515.348, df = 1, p < 0.0005), word count (t = -32.377, p < 0.0005), Flesch reading ease (t = 20.847 p < 0.0005), number of references (t = - 26.643, p < 0.0005), spelling and

For peer review only

grammatical mistakes using a US dictionary (t = -19.894, p < 0.0005), spelling and grammatical mistakes using a UK dictionary (t = -3.021, p = 0.003), first author ethnicity (chi square = 266.543, df = 5, p < 0.0005), and first author gender (chi squared = 181.058, df = 5, p < 0.0005).

Multivariate analysis

All variables above were used in a hierarchical, binary logistic regression with two blocks. The first included all variables except first author gender and ethnicity. The second block additionally contained first autho all r gender and ethnicity.

In the first block, which excluded author gender and ethnicity, there was a multivariate relationship between rapid response publication and: presence of references (OR 0.71 [95% CI 0.67–0.75], p < 0.0005), declaration of competing interests (OR 0.56 [95% CI 0.52–0.60], p < 0.0005), first author title containing the word “student” (OR 1.64 [95% CI 1.38–1.95], p < 0.0005), first author title containing the word “consultant” (OR 0.69 [95% CI 0.65–0.73], p < 0.0005), first author title containing the word “professor” (OR 0.75 [95% CI 0.70–0.81], p < 0.0005), first author title containing the word “senior” (OR 0.73 [95% CI 0.66–0.81], p < 0.0005), presence of multiple authors (OR 0.47 [95% CI 0.44–0.49], p < 0.0005), presence of a twitter handle (OR 0.69 [95% CI 0.60–0.79], p < 0.0005), word count being close to ideal (OR 0.54 [95% CI 0.51–0.57], p < 0.0005), Flesch reading ease being close to ideal (OR 0.80 [95% CI 0.76–0.85], p < 0.00055), word count (OR 1.00 [95% CI 1.00–1.00], p < 0.0005), Flesch reading ease (OR 1.00 [95% CI 1.00-1.00], p = 0.016), spelling and grammatical mistakes using a US dictionary (OR 1.00 [95% CI 1.00–1.00], p < 0.0005), spelling and grammatical mistakes using a UK dictionary (OR 1.00 [95% CI 0.99–1.00], p < 0.0005),and number of recent rapid response submissions (OR 1.00 [95% CI 1.00–1.00], p < 0.0005. Number of references was not significantly associated with rapid response publication in the multivariate model (OR 1.01 [95% CI 1.00-1.02], p = 0.290).

Commented [MZ3]: We added multivariate odds ratios for block one (i.e. with author ethnicity and gender excluded from the analysis) following comments from reviewer four.

First author gender and ethnicity remained statistically significant after accounting for measured confounders. In the second block, incorporating this information significantly improved the model (omnibus test of model coefficients, chi square = 4648.412, df = 43, p < 0.0005): in the second block, the pseudo R-squared value was 0.098, up from 0.88 in the first block.

Table 2 below shows the results of the second, complete logistic regression and odds ratios (OR) for each variable.

Table 2. Odds ratios with 95% confidence intervals and P values in multivariate analysis. “American Indian and Alaska Native” was removed from the ethnicity figures due to an absence of published letters from that ethnic group.

Commented [MZ4]: AIAN was spelled out following reviewer four's suggestion

Commented [MZ5]: These odds ratios have been rounded to the second decimal space, following the suggestion from reviewer four

Variable	Odds ratio (95% CI)	P value
Gender		
Male	–	< 0.0005
Female	0.90 (0.85–0.96)	0.002
Mostly male	0.97 (0.83–1.14)	0.727
Mostly female	0.98 (0.80–1.22)	0.882
Androgynous	1.02 (0.80–1.29)	0.901
Unknown	0.75 (0.69–0.81)	< 0.0005
Ethnicity		
White	–	< 0.0005
API	0.54 (0.49–0.59)	< 0.0005
Hispanic	0.51 (0.41–0.64)	< 0.0005
Black	0.74 (0.57–0.96)	0.023
Unknown	1.39 (0.99–1.96)	0.057
Word count	1.00 (1.00–1.00)	< 0.0005
Flesch reading ease	1.00 (1.00–1.00)	0.042

Has references	0.70 (0.67–0.74)	< 0.0005
Number of references	1.00 (0.99–1.01)	0.475
Author position		
Consultant	0.70 (0.65–0.74)	< 0.0005
Professor	0.74 (0.69–0.79)	< 0.0005
Senior	0.74 (0.67–0.81)	< 0.0005
Student	1.54 (1.30–1.83)	< 0.0005
Twitter handle present	0.69 (0.60–0.79)	< 0.0005
US spelling and grammar errors	1.00 (1.00–1.00)	0.001
UK spelling and grammar errors	0.99 (0.99–1.00)	< 0.0005
Multiple authors	0.44 (0.42–0.46)	< 0.0005
Competing interests declared	0.57 (0.53–0.61)	< 0.0005
Recent submission	1.00 (1.00– 1.00)	< 0.0005
Near ideal word count	0.54 (0.51–0.57)	< 0.0005
Near ideal Flesch reading ease	0.80 (0.76–0.85)	< 0.0005

Discussion:

Statement of principal findings

The estimated gender and ethnicity of first author names of BMJ rapid responses were predictive of publication, even when other features of the rapid response and the author were taken into account.

Strengths and weaknesses of the study

To the best of our knowledge, this is the largest ever analysis of a scientific corpus that looks for associations with publication rate. This was possible because of the open nature of BMJ rapid responses, and through automation at various stages, including data gathering and processing, which includes the use of validated machine learning algorithms for automatic ethnicity classification. This allowed us to analyse over 100,000 submissions, a feat which would not have been possible manually.

One of the largest weaknesses of this study is that it is only sensitive to associations. It is not possible to infer causality from this data, when the exact mechanisms for the observed discrepancy are not known. There could be other unmeasured factors accounting for the discrepancy in publication rates such as subtle differences in communication style, which have been posited to explain at least partly the ethnicity attainment gap in medical school clinical examinations (30). It is worth noting, however, that in clinical examinations it is male students that are consistently found to underperform relative to their female counterparts (31,32), while we found that female first authors were underrepresented.

Though the tool for classifying ethnicity from name has been validated on a global population, it was developed in the United States (US) and uses ethnicity categories that closely resemble those officially used within the US. This is not ideal for names in the UK, where different categories are defined officially. The categorisation of gender resulted in a fairly large proportion (17.3%) of authors with unknown gender, and they were less likely to be published. A similar tool for inferring gender from name, was shown to have an overall 93.8% accuracy in classifying author names in an analysis in the journal Science (33). This high accuracy is due to these techniques' ability to quantify their uncertainty; for example, if they believe there is a roughly equal chance of the name being male or female, it is classed as "androgynous". Only names which are very likely to be a specific gender are inferred as such.

Commented [MZ6]: Added more information about the gender inference tool based on reviewer two's comments.

A weakness in this work is that for all analysed letters, the corresponding rapid response had to be imputed using the protocol mentioned previously. A small minority of letters seem not to have been submitted as rapid responses, which may represent either direct publications from the editor or direct correspondences between a paper author and the editor. It is worth noting, however, that recently published letters link directly to the original rapid response that was submitted. This would allow future analysis to have ground truth data on which rapid responses were published and which were not.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Strengths and weaknesses in relation to other studies

A recent report by *Science* found no statistically significant evidence that their editorial process underrepresented female authors (34). In this report, gender was identified manually which limited their sample size to a small random selection of submissions. The gender disparity we observed is relatively subtle compared to that which we observed for ethnicity, and this might mean that larger sample sizes are needed to elucidate any discrepancies.

The under-representation of ethnicity has so far been less thoroughly studied in comparison to gender. Nonetheless, our findings are in line with published data that BME authors are underrepresented in published articles (10); however, our study is further able to identify that submissions from BME authors are less likely to be accepted for publication relative to similar submissions by their white counterparts.

Meaning of the study: possible mechanisms and implications

In our hierarchical logistic regression, gender and ethnicity explain a small amount of additional variance (0.01 increase in the pseudo R-squared) compared to the other features alone; however, this is considerable compared to a low pseudo R-squared baseline of 0.088.

Given the likely complexity of the selection process, it is unsurprising that the pseudo R-squared is low, as there are many unmeasured factors. Indeed, publication is determined by the expert opinion of the editor, on things that are impractical or impossible to quantify in a study like this, including clarity, style, and interest to the potential reader.

Nonetheless, one factor which may play a role is that of unconscious gender or racial bias. Implicit bias has been documented in clinical decision-making (35), medical school admissions (36), and selection of junior doctors (37). It is important to state, however, that the current study design does not provide causal evidence of bias, which would require a prospective experimental study design to fully account for other unmeasured factors.

Our results suggest that scientific journals should look for such discrepancies in all forms of submissions, including opinion pieces and research papers, which are not posted publicly. Such analysis should include looking at differences in submission and publication rates by author ethnicity as well as by author gender.

Unanswered questions and future research

This work highlights important associations in past data, however more research is necessary to draw concrete conclusions regarding the reasons for these associations. For example, communication style is not something that we accounted for in this work, and future studies can attempt to account for it through qualitative means to correct for it as a confounder. Additionally, studies have demonstrated unconscious gender biases in science (12), and unconscious racial biases in other areas (35–37), but far less research has studied unconscious biases in academia.

It is important to establish whether the discrepancies we found in BMJ letters to the editor are present in other journals, and for other scientific manuscript types such as original research. Though trends have been studied for published papers, quantifying the rate of acceptance is an invaluable way to eliminate the confounder that is number of submissions, and we hope that future research in this field can either be done by journals themselves, or by researchers in close collaboration with journals to ensure that this submission data is included in any analysis.

Data Sharing Statement: All data used in this manuscript is publicly available on: http://www.bmj.com. Large parts of the processed data used are available by emailing the corresponding author.

References:

1. Woolf K, Potts HWW, McManus IC. Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis. BMJ [Internet]. 2011 Mar 8 [cited 2017 Dec 16];342:d901.

Available from: http://www.ncbi.nlm.nih.gov/pubmed/21385802
2. General Medical Council. The state of medical education and practice in the UK. 2015 [cited 2019 Jul 20]; Available from: https://www.gmc-uk.org/-/media/documents/somep-2015_pdf-63501874.pdf
3. Appleby J. Ethnic pay gap among NHS doctors. BMJ [Internet]. 2018 Sep 5 [cited 2019 Feb 15];362:k3586. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30185418
4. Ginther DK, Schaffer WT, Schnell J, Masimore B, Liu F, Haak LL, et al. Race, ethnicity, and NIH research awards. Science [Internet]. 2011 Aug 19 [cited 2019 Jul 28];333(6045):1015–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21852498
5. Aamc. Analysis in Brief - July 2005: The changing representation of men and women in academic medicine [Internet]. 2005 [cited 2019 Mar 14]. Available from: www.aamc.org/data/aib
6. Rochon PA, Davidoff F, Levinson W. Women in Academic Medicine Leadership. Acad Med [Internet]. 2016 Aug [cited 2019 Mar 14];91(8):1053–6. Available from: http://insights.ovid.com/crossref?an=00001888-201608000-00014
7. Hopkins AL, Jawitz JW, McCarty C, Goldman A, Basu NB. Disparities in publication patterns by gender, race and ethnicity based on a survey of a random sample of authors. Scientometrics [Internet]. 2013 Aug 10 [cited 2019 Mar 13];96(2):515–34. Available from: http://link.springer.com/10.1007/s11192-012-0893-4
8. Sidhu R, Rajashekhar P, Lavin VL, Parry J, Attwood J, Holdcroft A, et al. The gender imbalance in academic medicine: a study of female authorship in the United Kingdom. J R Soc Med [Internet]. 2009 Aug 13 [cited 2019 Jul 20];102(8):337–42. Available from: http://journals.sagepub.com/doi/10.1258/jrsm.2009.080378
9. West JD, Jacquet J, King MM, Correll SJ, Bergstrom CT. The Role of Gender in Scholarly Authorship. Hadany L, editor. PLoS One [Internet]. 2013 Jul 22 [cited 2019 Jul 20];8(7):e66212. Available from: https://dx.plos.org/10.1371/journal.pone.0066212
10. Marschke G, Nunez A, Weinberg BA, Yu H. Last Place? The Intersection of Ethnicity, Gender, and Race in Biomedical Authorship. AEA Pap Proc [Internet]. 2018 [cited 2019 Jul 28];108:222–7. Available from: https://www.aeaweb.org/doi/10.1257/pandp.20181111
11. Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ, Handelsman J. Science faculty's subtle gender biases favor male students. Proc Natl Acad Sci U S A [Internet]. 2012 Oct 9 [cited 2019 Jul 20];109(41):16474–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22988126
12. Knobloch-Westerwick S, Glynn CJ, Huge M. The Matilda Effect in Science Communication. Sci Commun [Internet]. 2013 Oct 6 [cited 2019 Jul 20];35(5):603–25. Available from: http://journals.sagepub.com/doi/10.1177/1075547012472684
13. Ceci SJ, Williams WM. Understanding current causes of women's underrepresentation in science. Proc Natl Acad Sci U S A [Internet]. 2011 Feb 22 [cited 2019 Jul 20];108(8):3157–62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21300892
14. Boll C, Leppin J, Rossen A, Wolf A. Magnitude and Impact Factors of the Gender Pay Gap in EU Countries. 2016 [cited 2019 Jul 20]; Available from: www.fondazionebrodolini.it
15. Blau FD, Kahn LM. The Gender Wage Gap: Extent, Trends, and Explanations. J Econ Lit [Internet]. 2017 Sep [cited 2019 Jul 20];55(3):789–865. Available from: http://pubs.aeaweb.org/doi/10.1257/jel.20160995
16. Weaver AC, Wetterneck TB, Whelan CT, Hinami K. A matter of priorities? Exploring the persistent gender pay gap in hospital medicine. J Hosp Med [Internet]. 2015 Aug 1 [cited 2019 Jul 20];10(8):486–90. Available from: http://www.journalofhospitalmedicine.com/jhospmed/article/127833/priorities-and-gender-pay-gap
17. Ye J, Han S, Hu Y, Coskun B, Liu M, Qin H, et al. Nationality Classification Using Name Embeddings. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17 [Internet]. New York, New York, USA: ACM Press; 2017 [cited 2018 Apr 23]. p. 1897–906. Available from: http://dl.acm.org/citation.cfm?doid=3132847.3133008
18. Mateos P. A review of name-based ethnicity classification methods and their potential in population studies. Popul Space Place [Internet]. 2007 Jul 1 [cited 2018 Sep 28];13(4):243–63. Available from: http://doi.wiley.com/10.1002/psp.457
19. Banda Y, Kvale MN, Hoffmann TJ, Hesselson SE, Ranatunga D, Tang H, et al. Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. Genetics [Internet]. 2015 Aug [cited 2018 Sep 28];200(4):1285–95. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26092716
20. Lakha F, Gorman DR, Mateos P. Name analysis to classify populations by ethnicity in public health: Validation

For peer review only

of Onomap in Scotland. Public Health [Internet]. 2011 Oct 1 [cited 2018 Sep 28];125(10):688–96. Available from: https://www.sciencedirect.com/science/article/pii/S0033350611001508
21. Dukebody. Gender Guesser [Internet]. GitHub. 2018. Available from: https://github.com/lead-ratings/gender-guesser
22. Bansal S, Aggarwal C. textstat. 2018.
23. Myint. language-check. 2017.
24. Project Jupyter. Jupyter Notebook [Internet]. [cited 2019 Apr 16]. Available from: https://jupyter.org/index.html
25. kennethreitz. Requests: HTTP for Humans™ [Internet]. [cited 2019 Apr 16]. Available from: http://docs.python-requests.org/en/master/
26. Richardson L. Beautiful Soup [Internet]. [cited 2019 Apr 16]. Available from: https://www.crummy.com/software/BeautifulSoup/
27. Oliphant T. NumPy [Internet]. [cited 2019 Apr 16]. Available from: http://www.numpy.org/
28. McKinney W. Python Data Analysis Library [Internet]. [cited 2019 Apr 16]. Available from: https://pandas.pydata.org/
29. Cournapeau D. scikit-learn: machine learning in Python [Internet]. [cited 2019 Apr 16]. Available from: https://scikit-learn.org/stable/
30. Wass V, Roberts C, Hoogenboom R, Jones R, Van der Vleuten C. Effect of ethnicity on performance in a final objective structured clinical examination: qualitative and quantitative study. BMJ [Internet]. 2003 Apr 12 [cited 2019 Mar 14];326(7393):800–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12689978
31. Woolf K, Haq I, McManus IC, Higham J, Dacre J. Exploring the underperformance of male and minority ethnic medical students in first year clinical examinations. Adv Heal Sci Educ [Internet]. 2008 Dec 9 [cited 2019 Mar 14];13(5):607–16. Available from: http://link.springer.com/10.1007/s10459-007-9067-1
32. Lumb AB, Vail A. Comparison of academic, application form and social factors in predicting early performance on the medical course. Med Educ [Internet]. 2004 Sep 1 [cited 2019 Jul 28];38(9):1002–5. Available from: http://doi.wiley.com/10.1111/j.1365-2929.2004.01912.x
33. New tools for gender analysis | Sciencehound [Internet]. [cited 2020 Feb 14]. Available from: https://blogs.sciencemag.org/sciencehound/2019/01/03/new-tools-for-gender-analysis/
34. Berg J. Looking inward at gender issues. Science (80-) [Internet]. 2017 Jan 27 [cited 2018 Sep 25];355(6323):329–329. Available from: http://www.sciencemag.org/lookup/doi/10.1126/science.aam8109
35. Green AR, Carney DR, Pallin DJ, Ngo LH, Raymond KL, Iezzoni LI, et al. Implicit Bias among Physicians and its Prediction of Thrombolysis Decisions for Black and White Patients. J Gen Intern Med [Internet]. 2007 Aug 10 [cited 2019 Mar 14];22(9):1231–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17594129
36. Capers Q, Clinchot D, McDougle L, Greenwald AG. Implicit Racial Bias in Medical School Admissions. Acad Med [Internet]. 2017 Mar [cited 2019 Mar 14];92(3):365–9. Available from: http://insights.ovid.com/crossref?an=00001888-201703000-00032
37. Esmail A, Everington S. Racial discrimination against doctors from ethnic minorities. BMJ [Internet]. 1993 Mar 13 [cited 2019 Feb 12];306(6879):691–2. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8471921

For peer review only

BMJ Open

Gender and ethnic differences in publication of BMJ letters to the editor: an observational study using machine learning

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-037269.R1
Article Type:	Original research
Date Submitted by the Author:	29-Jul-2020
Complete List of Authors:	Zeina, Mohamad; Barts Health NHS Trust Balston, Alfred; Barts Health NHS Trust Banerjee, Amitava; University College London, Farr Institute of Health Informatics Research Woolf, Katherine; University College London, Research Department of Medical Education
Primary Subject Heading:	Medical publishing and peer review
Secondary Subject Heading:	Health informatics, Research methods, Public health, Medical management, Health services research
Keywords:	Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, EPIDEMIOLOGY, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Health economics < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, MEDICAL EDUCATION & TRAINING, STATISTICS & RESEARCH METHODS

SCHOLARONE™
Manuscripts

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

For peer review only

Gender and ethnic differences in publication of BMJ letters to the editor: an observational study using machine learning

Dr. Mohamad Zeina[1], Dr. Alfred Balston[1], Dr. Amitava Banerjee[2], Dr. Katherine Woolf[3]

[1] Foundation Year 1 Doctor, Barts Health NHS Trust, London, United Kingdom

[2] Associate Professor in Clinical Data Science and Honorary Consultant Cardiologist, Institute of Health Informatics, University College London, London, United Kingdom

[3] Associate Professor in Medical Education, Research Department of Medical Education, University College London Medical School, Royal Free Hospital, London, United Kingdom

Correspondence: Mohamad Zeina

Whipps Cross University Hospital

Barts Health NHS Trust,

London, E11 1NR

Tel: +44 7827353938

Email: mohamad.zeina@nhs.net

Copyright: The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd to permit this article (if accepted) to be published in BMJ editions and any other BMJPGL products and sublicences such use and exploit all subsidiary rights, as set out in our licence.

Competing interest statement: All authors have completed the Unified Competing Interest form and declare:

Dr. Zeina has nothing to disclose.

Dr. Balston has nothing to disclose.

Dr. Banerjee reports personal fees from Boehringer-Ingelheim, personal fees from Astra-Zeneca, personal fees from Novo-Nordisk, personal fees from Pfizer, outside the submitted work; and I am a Trustee of the South Asian Health Foundation, and a member of the Education Committee of the British Cardiovascular Society.

Dr. Woolf reports grants from National Institute for Health Research, non-financial support from Membership of the Royal Colleges of Physicians (UK) examination.

For peer review only

Guarantor: Mohamad Zeina

Contributorship statement:

- Mohamad Zeina: conceptualisation, design and coordination of project, lead in programming and statistical analysis, writing first draft.
- Alfred Balston: contribution to programming, statistical analysis, and writing first draft.
- Amitava Banerjee: aiding in conceptualisation, offering statistical advice, proofreading.
- Katherine Woolf: conceptualisation and design, offering statistical advice, proofreading.

The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted

Transparency declaration: Mohamad Zeina, affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Ethical approval: Ethical approval was not required.

Funding, sponsors and independence: This work is independent from funders, however other funding received by the authors is declared above in the competing interests statement.

Patient and Public Involvement: Patients were not involved in this study.

Dissemination statement: Dissemination not applicable as this study involved no participants or patient organisations.

For peer review only

Abstract

Objectives: To analyse the relationship between first author gender and ethnicity (estimated from first name and surname), and chance of publication of rapid responses in the BMJ. To analyse whether other features of the rapid response account for any gender or ethnic differences, including presence of multiple authors, declaration of conflicts of interests, presence of Twitter handle, word count, reading ease, spelling and grammatical mistakes, and presence of references.

Design: Retrospective observational study.

Setting: Website of the BMJ (BMJ.com).

Participants: Publicly available rapid responses submitted to BMJ.com between 1998 and 2018.

Main outcome measures: Publication of a rapid response as a letter to the editor in the BMJ.

Results: We analysed 113,265 rapid responses, of which 8,415 were published as letters to the editor (7.4%). Statistically significant univariate correlations were found between odds of publication and: first author estimated gender and ethnicity, multiple authors, declaration of conflicts of interest, presence of Twitter handle, word count, reading ease, spelling and grammatical mistakes, and presence of references. Multivariate analysis showed that first author estimated gender and ethnicity predicted publication after taking into account the other factors. Compared to white authors, black authors were 26% less likely to be published (OR 0.74, CI 0.57-0.96), Asian and Pacific Islander authors were 46% less likely to be published (OR 0.54, CI 0.49-0.59), and Hispanic authors were 49% less likely to be published (OR 0.51, CI 0.41-0.64). Female authors were 10% less likely to be published (OR 0.90, CI 0.85-0.96) than male authors.

Conclusion: Ethnic and gender differences in rapid response publication remained after accounting for a broad range of features, themselves all predictive of publication. This suggests that the reasons for the differences of these groups lies elsewhere.

Strengths and limitations of this study

- This study utilises corpus of publicly available data to analyse correlations between first author characteristics and chance of publication of letters to the editor in the BMJ.
- Multivariate analysis allowed us to account for a range of other features of submitted letters.
- To our knowledge, this is the largest ever analysis of a scientific corpus that looks at associations with publication rate.
- The nature of this data means that only associations can be inferred, and not causation.
- We highlight automated techniques that scientific journals can use to look for associations between ethnicity and gender in their own publication rates.

Introduction

Much has been written about the “attainment gap” or “differential attainment”; the observation that many fields exhibit discrepancies in achievement based on personal attributes such as gender and ethnicity. In medicine, for example, students from black and minority ethnic (BME) groups achieve poorer marks and are more likely to fail, on average, than their white counterparts (1). As they progress in their careers, BME doctors also more often fail their specialty training exams (1,2), earn a lower average salary than others at the same level of seniority (3), and are less likely to be awarded funding grants (4). There also remain discrepancies in the representation of women in medical leadership and faculty despite a long history of roughly equal proportions of male and female medical students (5,6).

Another specific area where the effects of gender have been studied thoroughly is academic publishing. A survey of 1065 authors from different backgrounds found that women were underrepresented in the scientific literature, along with certain ethnic minorities (7). In a group of high impact medical journals, including the British Medical

For peer review only

Journal (BMJ), the number of articles with female first authors has increased over time; however the gender balance of last authors, who are typically senior researchers or heads of departments, has not followed this trend (8). A larger scale analysis of 1.8 million scholarly articles indexed in JSTOR found female authors are poorly represented in the prestigious first and last author positions in the majority of academic disciplines (9). The same result was observed in an analysis of 21 million articles indexed in Medline which also found BME authors are similarly less likely to be in the last author position even when accounting for seniority. (10).

The cause of these gender and ethnic differences remains the subject of debate. Experimental studies have shown that identical submissions randomly assigned to have a male or female name are ranked differently depending on the gender of the applicant's name, favouring men (11,12). For example, applications for a laboratory manager position that were assigned a male name were rated as being significantly more competent and hireable by faculty members compared to identical applications assigned a female name (11). Similarly, a study presented graduate students with a sample of abstracts from an international conference, where abstracts were randomly shown to have male or female authors. The study found that abstracts presented as having female authors were deemed to have lower "Scientific Quality" (12). Both papers found that the gender of the reviewer did not affect how applicants were rated, and concluded that pervasive gender stereotypes create a subtle but significant bias against women (11,12). Indeed, the bias was reduced when reviewers' attitudes towards gender roles were taken into account, with higher support for gender equality being associated with higher ratings for female authors.

Some argue that gender disparities arise from men and women choosing different career paths (13). For example, women may opt to prioritise flexibility or take time out of their career to have children. However, studies that incorporate these factors into multivariate statistical models fail to fully account for discrepancies in pay (14,15). One such study found a $14,581 yearly salary difference between male and female hospital physicians in the United States which remained after accounting for differences in job satisfaction priorities between genders (16).

The underlying causes of these differences are likely to be complex and multifactorial, but identifying and characterising disparities in new specific situations might hint at potential solutions. These may be broadly applicable, especially because the causative issues are likely to compound each other. For example, lower average pay for women and BME doctors may be partly due to lower chances of scientific publication, especially in a work environment where publication in the scientific literature is important for attaining certain senior academic and leadership positions.

Though many studies mentioned here find group differences based on personal characteristics, they rarely have access to raw data from the journals that would be necessary to quantify publication rate. For example, a finding that women are under-represented in authorship of medical journal papers compared to in the medical workforce is not enough to draw conclusions about discrimination or bias, as it may be due to differences in priorities and the number of submissions sent. For a study to draw meaningful conclusions regarding discrepancies in acceptance rates, it must be able to quantify the percentage of submitted scientific works that are accepted, and this submission data is seldom released by scientific journals.

Letters in the BMJ are derived from rapid responses, which are available online freely and in their entirety, and therefore they may provide a valuable perspective for looking at this issue. Moreover, publication of rapid responses is of importance since letters to the editor carry PubMed identifiers (PMIDs) and thus discrepancies in their publication may have knock-on effects for jobs in academia where PubMed indexed publications play an important role in candidate selection.

We aimed to compare the corpus of available rapid responses with published letters to the editor to look for correlations between ethnicity, gender, and odds of publication.

For peer review only

Methods:

Data acquisition and processing

An automated script was used to download every BMJ.com online rapid response between 25th April 1998 and 23rd March 2018, as well as every letter to the editor that was published in the same timeframe.

To minimise the impact on BMJ servers, webpage requests were only sent every 15 seconds, and each request explicitly stated a full name and contact email address of the researcher carrying out the automated data collection, so that they could easily be contacted if the BMJ wished this collection to stop. Further, we only collected publicly available data that can be accessed without a login to a BMJ account.

Once collected, every available field from the rapid response was extracted. This included: title, title of article being responded to, body of text, first author name, first author title, other authors, date of submission, and presence of Twitter handle. Further processing with software packages mentioned below allowed us to look at a richer set of features including: word count, presence of references, number of references, Flesch reading ease (a measure of complexity of language, with a higher value meaning easier to read), number of spelling and grammatical mistakes, gender of first author, ethnicity of first author, and presence of multiple authors.

The position of the author was extracted by looking for the presence of each of the words "Consultant", "Professor", "Senior" and "Student" in the self-declared occupation field of submitted rapid responses, for example someone who had the word "Consultant" anywhere in their occupation field was classed "Consultant".

We did not expect a linear relationship between publication and word count or Flesch reading ease, because the most successful letters are likely to be long enough to offer a meaningful insight into the topic, but not too long as to be unsuitable for the short letter to the editor format. Thus we created two additional features from these, "Near ideal word count" and "Near ideal Flesch reading ease" to reflect whether a rapid response was within the 50% of rapid responses which are closest in word count and Flesch reading ease to the numbers which have historically been associated with higher rates of publication.

Some rapid responses (528, or 0.46%) could not be collected automatically due to errors in their formatting which prohibited their automated collection. These rapid responses were omitted from analysis. Regarding collected rapid responses, the absence of data was itself useful information (for example, the absence of 2nd authors was processed as there being no second 2nd authors) and so no analysed data point was considered missing.

As there is a lag between submission of a rapid response and publication of the response as a letter, we excluded all rapid responses that were within 66 days of our data collection window (i.e. submitted after 16th January 2018). This value was based on preliminary analysis that found a vast majority (80%) of letters were published within 66 days of the rapid response submission date.

Matching protocol

Although both rapid responses and published letters are available freely on BMJ.com, they are available on different parts of the website, and the vast majority of published letters do not link to the specific rapid response that was initially submitted. The task of finding out which rapid responses have been accepted is further complicated by the fact that that many editorial changes are made between the submission of a rapid response, and it being printed in the BMJ. Therefore, finding the corresponding rapid response for a letter is not as trivial as looking for a rapid response with identical text content.

To carry out this correspondence, a hierarchical matching protocol was used, which we summarise here. For each published letter, we search the corpus of rapid responses for those by the same first author. To make this possible, author names were standardised by removing middle names or initials. When a first author was only associated with a single rapid response, and a single letter to the editor, these were designated as the same submission. When the

For peer review only

author of a letter to the editor has submitted numerous rapid responses, one was chosen where the first 50 characters had the highest similarity with the letter to the editor.

If no rapid response could be found with the same first author as the published letter to the editor, author name was ignored and rapid responses submitted recently before publication of the letter to the editor were searched for one with the highest similarity in the first 50 characters. A subset of 600 matched rapid responses and letters was checked manually by AJB and MZ and found to be 85.3% accurate.

Classifying ethnicity and gender

Authors of rapid responses are not asked to disclose their ethnicity or gender, and the number of rapid responses involved was too great to individually contact the authors and ask this sensitive information. An automated method was used that could determine ethnicity of a name, for many tens of thousands of names, quickly and with little manual input. This took the shape of a previously published machine learning algorithm, *nameprism.com*, that has been trained to classify ethnicity on 74 million names, as well as being externally validated on datasets other than those which it was trained on. (17). To the best of our knowledge, it has demonstrated the highest classification accuracy of any publicly available tool for this task, with an F1 score of 0.795. We follow previous medical research that has used name to classify ethnicity (18,19), as well as a validation study that suggested name analysis is accurate enough to be used to aid health research (20).

This ethnicity classification tool was trained on a large, diverse set of names which the authors claim cover 90% of the world's names (17). It was developed in the US and so the six ethnicity categories used are American: White, Black, Asian and Pacific Islander (API), Hispanic, American Indian and Alaska Native (AIAN) and more than two races. Ethnicity was estimated using the first and last name of authors, which aims "to reduce errors when names are mixtures because of immigration or cross-nationality marriages". This tool is designed and tuned to infer on a world population, and so is well suited to a journal such as the BMJ with a worldwide authorship.

Gender of first author name was determined using a tool called *Gender Guesser* which utilises a database of approximately 40,000 common names and their corresponding gender (21). The first names of rapid response authors are checked against this database, and placed into one of the following categories: Male, Female, Mostly Male, Mostly Female, Androgynous (equal probabilities of being male or female) and unknown (not in the database). In an independent validation, on a manually labelled dataset of 7,076 names, it was compared to four other such gender inference tools, and was found to achieve "the lowest misclassification rate without parameter tuning for the entire dataset, introducing also the smallest gender bias" (22). We also validated the *Gender Guesser* tool on a public dataset of 29,872 names extracted from Wikipedia. The tool was able to infer gender for 82.76%. The names inferred as "male" were 99.2% accurate, and those inferred as "female" were 95.6% accurate. Overall, this tool was 98.4% accurate when detecting "male" or "female" names in our validation dataset .

The ethnicity and gender classification tools provide an estimated ethnicity and gender that, for the purpose of this study, is assumed to be analogous to the ethnicity and gender that a reader or reviewer would assign to an author.

Statistical analysis

Univariate associations between author and rapid response features, and publication was carried out by calculating chi square and t-test scores. Hierarchical binary logistic regression was used to look at the correlation between ethnicity and publication, taking into account other author and rapid response features.

Software used

Gender of first author names was classified using *Gender Guesser* (21). Ethnicity was classified using *nameprism.com* (17). Flesch reading ease score was calculated using an open source library called *textstat* (23). Spelling and grammatical mistakes were quantified using a tool called *language-check* (24).

For peer review only

All code was written in “Python 3” in the “Jupyter notebook” text editor (25). Data collection used an automated script, utilising the open source Python libraries “Requests” and “BeautifulSoup” (26,27). Further processing and data manipulation used the Python libraries “NumPy”, “Pandas” and “SciKit learn” (28–30). Statistical analysis was carried out in the IBM SPSS 25 package and in Python.

Results:

Baseline data

Analysis was perfomed on 113,265 rapid responses, of which 8,415 (7.4%) were published as letters to the editor. Of all submitted rapid responses, 83% had first authors with names classed as “white”; 62% of first authors were classed as “male”. See table 1 for baseline author and rapid response features. We also performed an analysis of the characteristics of submissions, broken down by inferred gender and ethnicity. These can be found in supplementary tables 1 and 2.

Table 1: Characteristics of published and unpublished rapid responses submitted to BMJ.com between 25th April 1998 and 23rd March 2018. API and AIAN stand for “Asian and Pacific Islander” and “American Indian and Alaska Native” respectively.
Plus-minus values are means ± SD. Percentages may not sum to 100% due to rounding.

Characteristic	All submissions (n=113265)	Published (n=8415)	Unpublished (n=104850)	**Statistical significance**
Author gender [number (%)]				Chi-sq=181, p < 0.0005
Male	70256 (62.0)	5636 (67.0)	64620 (61.6)	
Female	18592 (16.4)	1409 (16.7)	17183 (16.4)	
Mostly male	2434 (2.1)	171 (2.0)	2263 (2.2)	
Mostly female	1321 (1.2)	98 (1.2)	1223 (1.2)	
Androgynous	1021 (0.9)	82 (1.0)	939 (0.9)	
Unknown	19641 (17.3)	1019 (12.1)	18622 (17.8)	
Author ethnicity [number (%)]				Chi-sq=267, p < 0.0005
White	94077 (83.1)	7492 (89.0)	86585 (82.6)	
API	15759 (13.9)	726 (8.6)	15033 (14.3)	
Hispanic	1903 (1.7)	90 (1.1)	1813 (1.7)	
Black	1204 (1.1)	64 (0.8)	1140 (1.1)	
AIAN	2 (0.0)	0 (0.0)	2 (0.0)	
Unknown	320 (0.3)	43 (0.5)	277 (0.3)	
Word count	314 ± 318	410 ± 278	307 ± 319	t = -32.4, p < 0.0005
Flesch reading ease	50.3 ± 16.3	47.6 ± 12.1	50.5 ± 16.5	t = 20.8 p < 0.0005
Has references [number (%)]	40173 (35.5)	4445 (52.8)	35728 (34.1)	Chi-sq=1196, p < 0.0005
Number of references	1.3 ± 3.0	2.2 ± 3.3	1.3 ± 2.9	t = - 26.6, p < 0.0005
Author position [number (%)]				128, p < 0.0005
Consultant	16291 (14.4)	1592 (18.9)	14699 (14.0)	
Professor	9959 (8.8)	1110 (13.1)	8849 (8.4)	
Senior	4491 (4.0)	523 (6.2)	3968 (3.8)	
Student	3080 (2.7)	143 (1.7)	2937 (2.8)	
Other	79444 (70.1)	5047 (60.0)	74397 (71.0)	
Twitter handle present [number (%)]	1868 (0.2)	254 (0.3)	1614 (0.2)	Chi-sq=105, p < 0.0005

For peer review only

US spelling and grammar errors	27.0 ± 44.6	35.2 ± 39.2	26.3 ± 44.9	t = -19.9, p < 0.0005
UK spelling and grammar errors	8.9 ± 18.0	9.4 ± 15.3	8.9 ± 18.2	t = -3.0, p = 0.003
Multiple authors [number (%)]	19256 (17.0)	2914 (34.6)	16342 (15.6)	Chi-sq=2002, p <0.0005
Competing interests declared [number (%)]	6184 (5.5)	924 (11.0)	5260 (5.0)	Chi-sq=537, p < 0.0005

Univariate analysis

Univariate associations were analysed and included in table 1.

Multivariate analysis

All variables above were used in a hierarchical, binary logistic regression with two blocks. The first included all variables except first author gender and ethnicity. The second block additionally contained first author gender and ethnicity.

First author gender and ethnicity remained statistically significant after accounting for measured confounders. In the second block, incorporating this information significantly improved the model (omnibus test of model coefficients, chi square = 4648.412, df = 43, p < 0.0005): in the second block, the pseudo R-squared value was 0.098, up from 0.088 in the first block.

Table 2 below shows the results of the second, complete logistic regression and odds ratios (OR) for each variable.

Table 2. Odds ratios with 95% confidence intervals and P values in multivariate analysis. "American Indian and Alaska Native" was removed from the ethnicity figures due to an absence of published letters from that ethnic group. API stands for "Asian and Pacific Islander".

Variable	Block 1 - odds ratios (95% CI)	Block 1 - P value	Block 2 - odds ratio (95% CI)	Block 2 - P value
Gender				
Male	–	–	–	< 0.0005
Female	–	–	0.90 (0.85–0.96)	0.002
Mostly male	–	–	0.97 (0.83–1.14)	0.727
Mostly female	–	–	0.98 (0.80–1.22)	0.882
Androgynous	–	–	1.02 (0.80–1.29)	0.901
Unknown	–	–	0.75 (0.69–0.81)	< 0.0005
Ethnicity				
White	–	–	–	< 0.0005
API	–	–	0.54 (0.49–0.59)	< 0.0005
Hispanic	–	–	0.51 (0.41–0.64)	< 0.0005
Black	–	–	0.74 (0.57–0.96)	0.023
Unknown	–	–	1.39 (0.99–1.96)	0.057
Word count	1.00 (1.00 - 1.00)	< 0.0005	1.00 (1.00–1.00)	< 0.0005
Flesch reading ease	1.00 (1.00 - 1.00)	0.016	1.00 (1.00–1.00)	0.042
Has references	0.71 (0.67 - 0.75)	< 0.0005	0.70 (0.67–0.74)	< 0.0005
Number of references	1.01 (1.00 - 1.02)	0.290	1.00 (0.99–1.01)	0.475
Author position				
Consultant	0.69 (0.65 - 0.73)	< 0.0005	0.70 (0.65–0.74)	< 0.0005
Professor	0.75 (0.70 - 0.81)	< 0.0005	0.74 (0.69–0.79)	< 0.0005
Senior	0.73 (0.66 - 0.81)	< 0.0005	0.74 (0.67–0.81)	< 0.0005

For peer review only

Student	1.64 (1.38 - 1.95)	< 0.0005	1.54 (1.30–1.83)	< 0.0005
Twitter handle present	0.69 (0.60 - 0.79)	< 0.0005	0.69 (0.60–0.79)	< 0.0005
US spelling and grammar errors	1.00 (1.00 - 1.00)	< 0.0005	1.00 (1.00–1.00)	0.001
UK spelling and grammar errors	1.00 (0.99 - 1.00)	< 0.0005	0.99 (0.99–1.00)	< 0.0005
Multiple authors	0.47 (0.44 - 0.49)	< 0.0005	0.44 (0.42–0.46)	< 0.0005
Competing interests declared	0.56 (0.52 - 0.60)	< 0.0005	0.57 (0.53–0.61)	< 0.0005
Recent submission	1.00 (1.00 - 1.00)	< 0.0005	1.00 (1.00– 1.00)	< 0.0005
Near ideal word count	0.54 (0.51 - 0.57)	< 0.0005	0.54 (0.51–0.57)	< 0.0005
Near ideal Flesch reading ease	0.80 (0.76 - 0.85)	< 0.0005	0.80 (0.76–0.85)	< 0.0005

Discussion:

Statement of principal findings

The estimated gender and ethnicity of first author names of BMJ rapid responses were predictive of publication, even when other features of the rapid response and the author were taken into account.

Strengths and limitations of the study

To the best of our knowledge, this is the largest ever analysis of a scientific corpus that looks for associations with publication rate. This was possible because of the open nature of BMJ rapid responses, and through automation at various stages, including data gathering and processing, which includes the use of validated machine learning algorithms for automatic ethnicity classification. This allowed us to analyse over 100,000 submissions, a feat which would not have been possible manually.

One of the largest limitations of this study is that it is only sensitive to associations. It is not possible to infer causality from this data, when the exact mechanisms for the observed discrepancy are not known. There could be other unmeasured factors accounting for the discrepancy in publication rates such as subtle differences in communication style, which have been posited to explain at least partly the ethnicity attainment gap in medical school clinical examinations (31). It is worth noting, however, that in clinical examinations it is male students that are consistently found to underperform relative to their female counterparts (32,33), while we found that female first authors were underrepresented.

Though the tool for classifying ethnicity from name has been validated on a global population, it was developed in the United States (US) and uses ethnicity categories that closely resemble those officially used within the US. This is not ideal for names outside the US, where different categories are defined officially. The categorisation of gender resulted in a fairly large proportion (17.3%) of authors with unknown gender, and they were less likely to be published. Another similar tool for inferring gender from name, was shown to have an overall 93.8% accuracy in classifying author names in an analysis in the journal Science (34). This high accuracy is due to these techniques' ability to quantify their uncertainty; for example, if they believe there is a roughly equal chance of the name being male or female, it is classed as "androgynous". Only names which are very likely to be a specific gender are inferred as such.

Although there was an option to add location to the gender tool to determine the likely gender of a name in a specific country, this was not done for two reasons. Firstly, the location data extracted from rapid responses was highly heterogenous with some authors providing countries, cities, or institution names, or multiple addresses, without consistent spelling or abbreviations. Secondly, of the 45,376 names in the *Gender Guesser* dictionary, only 286 (0.6%) names are influenced sufficiently by location such that the estimate is changed from "male" to "female" or vice versa.

For peer review only

In this study, gender and ethnicity were estimated from the author's name, which provides a proxy for the gender and ethnicity that a reviewer would assign to an author given the same information. While assigning ethnicity and gender based on name may not always match the self-identified ethnicity and gender of the author, it is practical and necessary in this setting. A difficulty in classifying both ethnicity and gender into discreet categories is that the nuances of these complex social identities is lost, for example in the gender tool there is no categorisation for non-binary genders or transgender individuals, and limited provision for mixed race individuals in the ethnicity tool.

A limitation in this work is that for all analysed letters, the corresponding rapid response had to be imputed using the protocol mentioned previously. A small minority of letters seem not to have been submitted as rapid responses, which may represent either direct publications from the editor or direct correspondences between a paper author and the editor. It is worth noting, however, that recently published letters link directly to the original rapid response that was submitted. This would allow future analysis to have ground truth data on which rapid responses were published and which were not.

Strengths and limitations in relation to other studies

A recent report by *Science* found no statistically significant evidence that their editorial process underrepresented female authors (35). In this report, gender was identified manually which limited their sample size to a small random selection of submissions. The gender disparity we observed is relatively subtle compared to that which we observed for ethnicity, and this might mean that larger sample sizes are needed to elucidate any discrepancies.

The under-representation of ethnicity has so far been less thoroughly studied in comparison to gender. Nonetheless, our findings are in line with published data that BME authors are underrepresented in published articles (10); however, our study is further able to identify that submissions from BME authors are less likely to be accepted for publication relative to similar submissions by their white counterparts.

Implications

In our hierarchical logistic regression, gender and ethnicity explain a small amount of additional variance (0.01 increase in the pseudo R-squared) compared to the other features alone; however, this is considerable compared to a low pseudo R-squared baseline of 0.088.

Given the likely complexity of the selection process, it is unsurprising that the pseudo R-squared is low, as there are many unmeasured factors. Indeed, publication is determined by the expert opinion of the editor, on things that are impractical or impossible to quantify in a study like this, including clarity, style, and interest to the potential reader.

Nonetheless, one factor which may play a role is that of unconscious gender or racial bias. Implicit bias has been documented in clinical decision-making (36), medical school admissions (37), and selection of junior doctors (38). It is important to state, however, that the current study design does not provide causal evidence of bias, which would require a prospective experimental study design to fully account for other unmeasured factors.

Our results suggest that scientific journals should look for such discrepancies in all forms of submissions, including opinion pieces and research papers, which are not posted publicly. Such analysis should include looking at differences in submission and publication rates by author ethnicity as well as by author gender.

This study demonstrates the ability of data science and machine learning techniques to rapidly extract and analyse a large and complex dataset with relative ease. Without these techniques, this analysis would not have been possible. Being able to automate the process of feature extraction, including gender and ethnicity, opens avenues for further observational studies of open access data. It has also opened countless possibilities in medicine across the entire patient journey, from triage and improving attendance, to automated disease diagnosis, prognostication, management, and even the discovery or repurposing of new medications (39–45).

For peer review only

Unanswered questions and future research

This work highlights important associations in past data, however more research is necessary to draw concrete conclusions regarding the reasons for these associations. For example, other confounders might be considered, including communication style, field of study of the rapid response author and of the article being responded to, and the locations of institutes of submitted pieces. It may also be interesting to see how these discrepancies change over time, if at all. Additionally, studies have demonstrated unconscious gender biases in science (12), and unconscious racial biases in other areas (36–38), but far less research has studied unconscious biases in academia.

It is important to establish whether the discrepancies we found in BMJ letters to the editor are present in other journals, and for other scientific manuscript types such as original research. Though trends have been studied for published papers, quantifying the rate of acceptance is an invaluable way to eliminate the confounder that is number of submissions, and we hope that future research in this field can either be done by journals themselves, or by researchers in close collaboration with journals to ensure that this submission data is included in any analysis.

Conclusion

A number of variables were identified that correlated with the acceptance rate of rapid responses. Discrepancies in the publication rates between genders and ethnicities remained significant after accounting for other factors. The cause of these discrepancies is unclear and may in part be explicable by implicit bias. Regardless of the cause, it is evident that female and BME voices are underrepresented, and efforts should be made to identify these causes and rectify them.

Data Sharing Statement: All data used in this manuscript is publicly available on: http://www.bmj.com. Large parts of the processed data used are available by emailing the corresponding author.

References:

1. Woolf K, Potts HWW, McManus IC. Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis. BMJ [Internet]. 2011 Mar 8 [cited 2017 Dec 16];342:d901. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21385802
2. General Medical Council. The state of medical education and practice in the UK. 2015 [cited 2019 Jul 20]; Available from: https://www.gmc-uk.org/-/media/documents/somep-2015_pdf-63501874.pdf
3. Appleby J. Ethnic pay gap among NHS doctors. BMJ [Internet]. 2018 Sep 5 [cited 2019 Feb 15];362:k3586. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30185418
4. Ginther DK, Schaffer WT, Schnell J, Masimore B, Liu F, Haak LL, et al. Race, ethnicity, and NIH research awards. Science [Internet]. 2011 Aug 19 [cited 2019 Jul 28];333(6045):1015–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21852498
5. Aamc. Analysis in Brief - July 2005: The changing representation of men and women in academic medicine [Internet]. 2005 [cited 2019 Mar 14]. Available from: www.aamc.org/data/aib
6. Rochon PA, Davidoff F, Levinson W. Women in Academic Medicine Leadership. Acad Med [Internet]. 2016 Aug [cited 2019 Mar 14];91(8):1053–6. Available from: http://insights.ovid.com/crossref?an=00001888-201608000-00014
7. Hopkins AL, Jawitz JW, McCarty C, Goldman A, Basu NB. Disparities in publication patterns by gender, race and ethnicity based on a survey of a random sample of authors. Scientometrics [Internet]. 2013 Aug 10 [cited 2019 Mar 13];96(2):515–34. Available from: http://link.springer.com/10.1007/s11192-012-0893-4
8. Sidhu R, Rajashekhar P, Lavin VL, Parry J, Attwood J, Holdcroft A, et al. The gender imbalance in academic medicine: a study of female authorship in the United Kingdom. J R Soc Med [Internet]. 2009 Aug 13 [cited 2019 Jul 20];102(8):337–42. Available from: http://journals.sagepub.com/doi/10.1258/jrsm.2009.080378
9. West JD, Jacquet J, King MM, Correll SJ, Bergstrom CT. The Role of Gender in Scholarly Authorship. Hadany L, editor. PLoS One [Internet]. 2013 Jul 22 [cited 2019 Jul 20];8(7):e66212. Available from: https://dx.plos.org/10.1371/journal.pone.0066212
10. Marschke G, Nunez A, Weinberg BA, Yu H. Last Place? The Intersection of Ethnicity, Gender, and Race in

For peer review only

Biomedical Authorship. AEA Pap Proc [Internet]. 2018 [cited 2019 Jul 28];108:222–7. Available from: https://www.aeaweb.org/doi/10.1257/pandp.20181111
11. Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ, Handelsman J. Science faculty's subtle gender biases favor male students. Proc Natl Acad Sci U S A [Internet]. 2012 Oct 9 [cited 2019 Jul 20];109(41):16474–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22988126
12. Knobloch-Westerwick S, Glynn CJ, Huge M. The Matilda Effect in Science Communication. Sci Commun [Internet]. 2013 Oct 6 [cited 2019 Jul 20];35(5):603–25. Available from: http://journals.sagepub.com/doi/10.1177/1075547012472684
13. Ceci SJ, Williams WM. Understanding current causes of women's underrepresentation in science. Proc Natl Acad Sci U S A [Internet]. 2011 Feb 22 [cited 2019 Jul 20];108(8):3157–62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21300892
14. Boll C, Leppin J, Rossen A, Wolf A. Magnitude and Impact Factors of the Gender Pay Gap in EU Countries. 2016 [cited 2019 Jul 20]; Available from: www.fondazionebrodolini.it
15. Blau FD, Kahn LM. The Gender Wage Gap: Extent, Trends, and Explanations. J Econ Lit [Internet]. 2017 Sep [cited 2019 Jul 20];55(3):789–865. Available from: http://pubs.aeaweb.org/doi/10.1257/jel.20160995
16. Weaver AC, Wetterneck TB, Whelan CT, Hinami K. A matter of priorities? Exploring the persistent gender pay gap in hospital medicine. J Hosp Med [Internet]. 2015 Aug 1 [cited 2019 Jul 20];10(8):486–90. Available from: http://www.journalofhospitalmedicine.com/jhospmed/article/127833/priorities-and-gender-pay-gap
17. Ye J, Han S, Hu Y, Coskun B, Liu M, Qin H, et al. Nationality Classification Using Name Embeddings. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17 [Internet]. New York, New York, USA: ACM Press; 2017 [cited 2018 Apr 23]. p. 1897–906. Available from: http://dl.acm.org/citation.cfm?doid=3132847.3133008
18. Mateos P. A review of name-based ethnicity classification methods and their potential in population studies. Popul Space Place [Internet]. 2007 Jul 1 [cited 2018 Sep 28];13(4):243–63. Available from: http://doi.wiley.com/10.1002/psp.457
19. Banda Y, Kvale MN, Hoffmann TJ, Hesselson SE, Ranatunga D, Tang H, et al. Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. Genetics [Internet]. 2015 Aug [cited 2018 Sep 28];200(4):1285–95. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26092716
20. Lakha F, Gorman DR, Mateos P. Name analysis to classify populations by ethnicity in public health: Validation of Onomap in Scotland. Public Health [Internet]. 2011 Oct 1 [cited 2018 Sep 28];125(10):688–96. Available from: https://www.sciencedirect.com/science/article/pii/S0033350611001508
21. Dukebody. Gender Guesser [Internet]. GitHub. 2018. Available from: https://github.com/lead-ratings/gender-guesser
22. Santamaría L, Mihaljević H. Comparison and benchmark of name-to-gender inference services.
23. Bansal S, Aggarwal C. textstat. 2018.
24. Myint. language-check. 2017.
25. Project Jupyter. Jupyter Notebook [Internet]. [cited 2019 Apr 16]. Available from: https://jupyter.org/index.html
26. kennethreitz. Requests: HTTP for Humans™ [Internet]. [cited 2019 Apr 16]. Available from: http://docs.python-requests.org/en/master/
27. Richardson L. Beautiful Soup [Internet]. [cited 2019 Apr 16]. Available from: https://www.crummy.com/software/BeautifulSoup/
28. Oliphant T. NumPy [Internet]. [cited 2019 Apr 16]. Available from: http://www.numpy.org/
29. McKinney W. Python Data Analysis Library [Internet]. [cited 2019 Apr 16]. Available from: https://pandas.pydata.org/
30. Cournapeau D. scikit-learn: machine learning in Python [Internet]. [cited 2019 Apr 16]. Available from: https://scikit-learn.org/stable/
31. Wass V, Roberts C, Hoogenboom R, Jones R, Van der Vleuten C. Effect of ethnicity on performance in a final objective structured clinical examination: qualitative and quantitative study. BMJ [Internet]. 2003 Apr 12 [cited 2019 Mar 14];326(7393):800–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12689978
32. Woolf K, Haq I, McManus IC, Higham J, Dacre J. Exploring the underperformance of male and minority ethnic medical students in first year clinical examinations. Adv Heal Sci Educ [Internet]. 2008 Dec 9 [cited 2019 Mar

For peer review only

14];13(5):607–16. Available from: http://link.springer.com/10.1007/s10459-007-9067-1

33. Lumb AB, Vail A. Comparison of academic, application form and social factors in predicting early performance on the medical course. Med Educ [Internet]. 2004 Sep 1 [cited 2019 Jul 28];38(9):1002–5. Available from: http://doi.wiley.com/10.1111/j.1365-2929.2004.01912.x
34. New tools for gender analysis | Sciencehound [Internet]. [cited 2020 Feb 14]. Available from: https://blogs.sciencemag.org/sciencehound/2019/01/03/new-tools-for-gender-analysis/
35. Berg J. Looking inward at gender issues. Science (80-) [Internet]. 2017 Jan 27 [cited 2018 Sep 25];355(6323):329–329. Available from: http://www.sciencemag.org/lookup/doi/10.1126/science.aam8109
36. Green AR, Carney DR, Pallin DJ, Ngo LH, Raymond KL, Iezzoni LI, et al. Implicit Bias among Physicians and its Prediction of Thrombolysis Decisions for Black and White Patients. J Gen Intern Med [Internet]. 2007 Aug 10 [cited 2019 Mar 14];22(9):1231–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17594129
37. Capers Q, Clinchot D, McDougle L, Greenwald AG. Implicit Racial Bias in Medical School Admissions. Acad Med [Internet]. 2017 Mar [cited 2019 Mar 14];92(3):365–9. Available from: http://insights.ovid.com/crossref?an=00001888-201703000-00032
38. Esmail A, Everington S. Racial discrimination against doctors from ethnic minorities. BMJ [Internet]. 1993 Mar 13 [cited 2019 Feb 12];306(6879):691–2. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8471921
39. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med [Internet]. 2018 Sep 13 [cited 2019 Sep 2];24(9):1342–50. Available from: http://www.nature.com/articles/s41591-018-0107-6
40. Nelson A, Herron D, Rees G, Nachev P. Predicting scheduled hospital attendance with artificial intelligence. npj Digit Med [Internet]. 2019 Dec 12 [cited 2020 Jul 26];2(1):1–7. Available from: https://doi.org/10.1038/s41746-019-0103-3
41. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 2017 Nov 14 [cited 2018 Dec 15]; Available from: http://arxiv.org/abs/1711.05225
42. Alaa AM, van der Schaar M. Prognostication and Risk Factors for Cystic Fibrosis via Automated Machine Learning. Sci Rep [Internet]. 2018 Dec 1 [cited 2020 Jul 26];8(1):11242. Available from: www.nature.com/scientificreports
43. Longmire M. US8548828B1 - Method, process and system for disease management using machine learning process and electronic media - Google Patents [Internet]. [cited 2020 Jul 26]. Available from: https://patents.google.com/patent/US8548828B1/en
44. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A Deep Learning Approach to Antibiotic Discovery. Cell. 2020 Feb 20;180(4):688-702.e13.
45. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Vol. 20, The Lancet Oncology. Lancet Publishing Group; 2019. p. e262–73.

For peer review only

Supplementary Table 1: Characteristics of rapid responses submitted to BMJ.com between 25th April 1998 and 23rd March 2018 broken down by gender. API and AIAN stand for "Asian and Pacific Islander" and "American Indian and Alaska Native" respectively. Plus-minus values are means ± SD. Percentages may not sum to 100% due to rounding.

Characteristic	Male (n=70256)	Female (n=18592)	Mostly male (n=2434)	Mostly female (n=1321)	Androgynous (n=1021)	Unknown (n=19641)
Author ethnicity [number (%)]						
White	63562 (90.5)	16567 (89.1)	1635 (67.2)	1145 (86.7)	449 (44.0)	10719 (54.6)
API	4765 (6.7)	1686 (9.1)	740 (30.4)	155 (11.7)	548 (53.7)	7865 (40.0)
Hispanic	1400 (2.0)	189 (1.0)	38 (1.6)	17 (1.3)	1 (0.1)	258 (1.3)
Black	325 (0.5)	99 (0.5)	15 (0.6)	2 (0.2)	22 (2.2)	741 (3.8)
AIAN	1 (0.0)	1 (0.0)	0 (0.0)	0 (0.0)	0 (0)	0 (0.0)
Unknown	203 (0.3)	50 (0.3)	6 (0.2)	2 (0.2)	1 (0.1)	58 (0.3)
Word count	314 ± 324	330 ± 337	310 ± 264	322 ± 327	360 ± 319	297 ± 280
Flesch reading ease	50.0 ± 15.9	50.4 ± 16.1	50.3 ± 13.6	52.0 ± 14.3	48.2 ± 16.9	51.1 ± 18.0
Has references [number (%)]	25654 (36.5)	6087 (32.7)	853 (35.0)	405 (30.7)	463 (45.3)	6711 (34.2)
Number of references	1.4 ± 3.0	1.2 ± 3.1	1.2 ± 2.5	1.0 ± 2.3	2.3 ± 4.1	1.2 ± 2.6
Author position [number (%)]						
Consultant	11408 (16.2)	1303 (7.0)	345 (14.2)	138 (10.4)	150 (14.7)	2947 (15.0)
Professor	6570 (9.4)	1001 (5.4)	252 (10.4)	110 (8.3)	138 (13.5)	1888 (9.6)
Senior	2694 (3.8)	847 (4.6)	78 (3.2)	79 (6.0)	42 (4.1)	751 (3.8)
Student	1483 (2.1)	796 (4.3)	47 (1.9)	82 (6.2)	42 (4.1)	630 (3.2)
Other	48101 (68.5)	14645 (78.8)	1712 (70.3)	912 (69.0)	649 (63.6)	13425 (68.4)
Twitter handle present [number (%)]	1118 (1.6)	345 (1.9)	50 (2.1)	23 (1.7)	22 (2.2)	310 (1.6)
US spelling and grammar errors	27.1 ± 47.0	27.7 ± 44.9	26.3 ± 25.8	25.9 ± 37.9	32.8 ± 42.7	25.7 ± 35.8
UK spelling and grammar errors	8.9 ± 17.6	9.2 ± 20.1	9.0 ± 20.2	9.7 ± 15.4	8.0 ± 15.8	8.7 ± 17.2
Multiple authors [number (%)]	10436 (14.9)	4386 (23.6)	401 (16.5)	227 (17.2)	299 (29.3)	3507 (17.9)
Competing interests declared [number (%)]	3902 (5.6)	1011 (5.4)	99 (4.1)	55 (4.2)	88 (8.6)	1029 (5.2)

For peer review only

Supplementary Table 2: Characteristics of rapid responses submitted to BMJ.com between 25th April 1998 and 23rd March 2018 broken down by ethnicity. API and AIAN stand for "Asian and Pacific Islander" and "American Indian and Alaska Native" respectively. Plus-minus values are means ± SD. Percentages may not sum to 100% due to rounding.

Characteristic	White (n=94077)	API (n=15759)	Hispanic (n=1903)	Black (n=1203)	AIAN (n=2)	Unknown (n=320)
Author gender [number (%)]						
Male	63562 (67.6)	4765 (30.2)	1400 (73.6)	325 (27.0)	1 (50)	203 (63.4)
Female	16567 (17.6)	1686 (10.7)	189 (9.9)	99 (8.2)	1 (50)	50 (15.6)
Mostly male	1635 (1.7)	740 (4.7)	38 (2.0)	15(1.2)	0	6 (1.9)
Mostly female	1145 (1.2)	155 (1.0)	17 (0.9)	2 (0.2)	0	2 (0.6)
Androgynous	449 (0.5)	548 (3.5)	1 (0.1)	22 (1.8)	0	1 (0.3)
Unknown	10719 (11.4)	7865 (49.9)	258 (13.6)	741 (61.5)	0	58 (18.1)
Word count	318 ± 328	295 ± 259	317 ± 267	316 ± 279	157 ± 2	310 ± 247
Flesch reading ease	50.3 ± 16.4	50.7 ± 15.7	49.7 ± 14.4	47.7 ± 16.7	18.4 ± 4.5	47.2 ± 15.6
Has references [number (%)]	32887 (35.0)	5990 (38.0)	769 (40.4)	391 (32.5)	0 (0)	136 (42.5)
Number of references	1.3 ± 3.0	1.4 ± 2.7	1.6 ± 2.8	1.2 ± 2.8	0.0 ± 0.0	1.7 ± 2.9
Author position [number (%)]						
Consultant	13662 (14.5)	2276 (14.4)	82 (4.3)	222 (18.5)	0 (0.0)	49 (15.3)
Professor	7821 (8.3)	1715 (10.9)	280 (14.7)	114 (9.5)	0 (0.0)	29 (9.1)
Senior	3691 (3.9)	707 (4.5)	37 (1.9)	50 (4.2)	0 (0.0)	6 (1.9)
Student	2231 (2.4)	727 (4.6)	69 (3.6)	37 (3.1)	0 (0.0)	16 (5.0)
Other	66672 (70.9)	10334 (65.6)	1435 (75.4)	781 (64.9)	2 (100)	220 (68.8)
Twitter handle present [number (%)]	1547 (1.6)	283 (1.8)	27 (1.4)	11 (0.9)	0 (0.0)	0 (0.0)
US spelling and grammar errors	27.2 ± 46.2	25.5 ± 35.7	29.7 ± 34.6	26.3 ± 32.1	13.0 ± 4.2	25.5 ± 29.3
UK spelling and grammar errors	9.1 ± 18.3	7.7 ± 17.5	7.9 ± 11.7	9.5 ± 15.2	7.0 ± 9.9	5.9 ± 11.2
Multiple authors [number (%)]	14580 (15.5)	3865 (24.5)	531 (27.9)	207 (17.2)	0 (0)	73 (22.8)
Competing interests declared [number (%)]	5047 (5.4)	939 (6.0)	82 (4.3)	37 (3.1)	0 (0)	79 (24.7)

STROBE Statement—checklist of items that should be included in reports of observational studies

	Item No	Recommendation	Done
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	Yes
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	Yes
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	Yes
Objectives	3	State specific objectives, including any prespecified hypotheses	Yes
Methods			
Study design	4	Present key elements of study design early in the paper	Yes
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	Yes
Participants	6	(*a*) *Cohort study*—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study*—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study*—Give the eligibility criteria, and the sources and methods of selection of participants	N/A
		(*b*) *Cohort study*—For matched studies, give matching criteria and number of exposed and unexposed *Case-control study*—For matched studies, give matching criteria and the number of controls per case	N/A
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	Yes
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	Yes
Bias	9	Describe any efforts to address potential sources of bias	Yes
Study size	10	Explain how the study size was arrived at	Yes
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	Yes
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	Yes
		(*b*) Describe any methods used to examine subgroups and interactions	Yes
		(*c*) Explain how missing data were addressed	Yes
		(*d*) *Cohort study*—If applicable, explain how loss to follow-up was addressed *Case-control study*—If applicable, explain how matching of cases and controls was addressed *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy	N/A
		(*e*) Describe any sensitivity analyses	N/A

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Results			Done
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	N/A
		(b) Give reasons for non-participation at each stage	N/A
		(c) Consider use of a flow diagram	N/A
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	Yes
		(b) Indicate number of participants with missing data for each variable of interest	Yes
		(c) *Cohort study*—Summarise follow-up time (eg, average and total amount)	N/A
Outcome data	15*	*Cohort study*—Report numbers of outcome events or summary measures over time	N/A
		Case-control study—Report numbers in each exposure category, or summary measures of exposure	N/A
		Cross-sectional study—Report numbers of outcome events or summary measures	N/A
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	Yes
		(*b*) Report category boundaries when continuous variables were categorized	Yes
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	N/A
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	Yes
Discussion			
Key results	18	Summarise key results with reference to study objectives	Yes
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	Yes
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	Yes
Generalisability	21	Discuss the generalisability (external validity) of the study results	Yes
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	Yes

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml